Povídání na téma

SUPERPOČÍTAČE DNES A ZÍTRA
(aneb krátký náhled na SC)




IT4Innovations
národní superpočítačové centrum

29. 10. 2015 Filip Staněk

Osnova

- Co jsou to „Superpočítače“?
- Výkon SC
- Architektura
- Software
- Algoritmy
- IT4Innovations
- Odkazy na další informace

IT4Innovations#
národní01#$%@&0
superpočítačové
centrum$@00&1@&

Co je to vlastně SC?

Výpočetní systém, který určuje hranici maximálního možného výpočetního výkonu...

...v dané době samozřejmě.

Jak poznat onu maximální hranici v konkrétní době?

www.top500.org

IT4Innovations#
národní01#$%@&0
superpočítačové
centrum$@00&1@&

Jak se stanovuje výkon (TOP500)?

FLOP/s – počet operací s plovoucí čárkou za 1 s.
(konkrétně se jedná o 64bitové číslo s plovoucí čárkou a operaci sčítání nebo násobení)
Dnes se používá spíše TFLOP/s (10^{12} FLOP/s).

Rpeak – maximální teoretický výkon pro jedno jádro: PO*FR
(PO – počet operací za cyklus, FR – frekvence CPU)

Rmax – výkon naměřený v High Performance Linpack (HPL) benchmarku

Linpack – program na měření výpočetního výkonu, pomocí řešení husté soustavy lineárních rovnic

Jiné druhy požadavků na výkon

- Mnoho úloh spadá do problematiky řešení úloh grafů
- Někdy je prioritní cena za výsledek (ne čas)
- Mnohé úlohy reprezentují řídké matice

  Graph500.org, Green500.org, HPCG-benchmark.org

- Mnohé úlohy nepotřebují vysokou míru paralelizace – jde spíše o distribuované výpočty

  Cloud computing, BOINC, Bitcoins

TOP500 (červen 2015)

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	Swiss National Supercomputing Centre (CSCS) Switzerland	**Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Cray Inc.	115,984	6,271.0	7,788.9	2,325
7	King Abdullah University of Science and Technology Saudi Arabia	**Shaheen II** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	196,608	5,537.0	7,235.2	2,834
8	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P Dell	462,462	5,168.1	8,520.1	4,510
9	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	458,752	5,008.9	5,872.0	2,301
10	DOE/NNSA/LLNL United States	**Vulcan** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393,216	4,293.3	5,033.2	1,972

PRACE

IT4Innovations#
národní01#$%@&0
superpočítačové
centrum$@00&1@&

Architektura

- Architektura clusteru, uzly jsou víceprocesorové
- Používaná CPU – Intel x86_64 (často s akcelerátory), IBM Power BQC (BlueGene/Q), SPARC64 (Kcomp.)
- Výpočetní síť – TH Express-2 (Tianhe), Aries (CRAY), BG/Q, Tofu (Kcomp.), Infiniband

Tianhe-2 (cluster)

- Cluster, Intel IvyBridge (minulá generace), vlastní interconnect TH Express-2, Intel Xeon Phi koprocessory 31S1P (8GB RAM)
- Velká velikost (16k uzlů, 24C+2Phi/uzel, 88GB RAM)
- Vysoký výpočetní výkon (alespoň v TOP500)
- Mizerná efektivita (61%)
- Brutální pořizovací i provozní náklady (24MW včetně chlazení)

TITAN (Cray)

- Cluster, AMD Opteron (minulá generace), vlastní interconnect Gemini (minulá generace), NVIDIA K20x (Kepler, 6GB)
- Velká velikost (18,6k uzlů, 16C+1NVIDIA/uzel, 32GB RAM)
- Velký výpočetní výkon (alespoň v TOP500)
- Nízká efektivita (64,8%)
- Vysoké pořizovací i provozní náklady (8.2MW, chlazení až 23.2MW)

Sequoia (IBM BlueGene/Q)

- Vysoce paralelní specifický cluster, vlastní výkonný interconnect, spousta „pomalejších“ jader (1.6GHz)
- Brutální velikost (98.3k uzlů, 16C/uzel, 16GB RAM)
- Obecně velký výpočetní výkon pro dobře škálující SW
- Vysoká efektivita (85%)
- Lepší provozní náklady (spotřeba 7.8MW), chlazení vodou, nejvýše postavený „general purpose CPU“ v Green500

K computer (Fujitsu)

- Paralelní specifický cluster, vlastní výkonný interconnect, specifická jádra (SPARC64)
- Velký počet jader (82.9k uzlů, 8C/uzel, 16GB RAM)
- Velký výpočetní výkon (alespoň v TOP500)
- Vysoká efektivita (93%)
- Vysoká cena, ne tak dobrá efektivita napájení/chlazení

Důležité prvky výkonu

- CPU – dlouhé registry pro výpočty (AVX, AVX2, AVX3/512), FMA instrukce
- Rychlý přístup do paměti (DDR4, GDDR5 u GPU/Phi)
- Rychlý interconnect (včetně rychlého přístupu k CPU/RAM) s dobrou topologií

- Akcelerátory…

Akcelerátory

- GPU – hlavně NVIDIA (CUDA ecosystém), ADM s OpenCL spíše zaostává (může se změnit i díky Phi)
- Speciální programování (jak jazyk, tak postupy), vhodné pouze pro určité úlohy, limitace dostupné paměti (množství i přístup)

- Intel Xeon Phi – programování podobné OpenMP (offload) nebo nativní (jako X86 CPU), zatím kratší doba na trhu (menší podpora), taktéž limitace s lokální pamětí

Budoucí technologie

- Vývoj klasických CPU (Intel integruje vlastnosti Phi), AMD integruje vlastnosti svých GPU (ZEN), IBM otevírá Power (OpenPOWER)
- ARM CPU začínají podporovat 64bit, interconnecty a výkon již není „zcela nezajímavý“
- GPU jako hlavní zdroj výkonu (Volta, NVLink)
- Xeon Phi jak standalone uzel (KNL)
- GPU i Xeon Phi s HBM pamětmi (propustnost i latence)

Budoucí technologie II

- Mellanox s EDR Infinband
- Intel s OMNI-PATH
- ATOS s BXI

- SSD pro disková úložiště jako primární prvky

- Intel & Micron 3D Xpoint paměti
- Memristory? Silicon Photonics? Kvantové počítače?

Software

- Lehký OS (minimalizace jitteru/přerušení)
- Ovladače a knihovny (obcházející OS) pro interconnect
- Uživatelské knihovny pro komunikaci (MPI, GASPI)
- Kvalitní překladače (potřeba optimalizací)
- Analyzátory a debuggery

- Plánovače/manažery zdrojů
- Paralelní souborové systémy

Programovací modely pro SC

- Vektorizace
- Paralelizace (OpenMP, MPI, hybridní)
- Minimalizace I/O (čtení a zápis počas běhu výpočtu)
- Efektivní komunikace (communication latency hiding)

- Potřeba mít dobrou (optimálně lineární) silnou i slabou škálovatelnost.

IT4Innovations

- Národní superpočítačové centrum v ČR (Ostrava, VŠB-TUO), spolu s CESNET národní E-Infrastruktura
- 8 vlastních výzkumných programů
- Zástupce v PRACE za ČR, řešitel všech Impl. Fází PRACE
- Intel Parallel Computing CENTER
- Řešitel 4 H2020 projektů a 2 EC FP7 projektů
- Spolupráce na projektech s ESA
- Řešitel projektů GAČR, TAČR

IT4Innovations#
národní01#$%@&0
superpočítačové
centrum$@00&1@&

IT4I Infrastruktura

- Pokrývá samostatná prezentace, obsahem je:

- Vlastní datový sál, včetně energocentra a chlazení
- Dva superpočítače, výkonnější na 40. místě TOP500

- Přístup k testování nových technologií (Intel, Melllanox)

IT4Innovations#
národní01#$%@&0
superpočítačové
centrum$@00&1@&

Zdroje informací

www.top500.org

www.graph500.org

www.green500.org

www.hpcg-benchmark.org

www.it4i.cz

www.hpcwire.com

www.wikipedia.org

www.google.com

Děkuji za pozornost.

Otázky?

www.it4i.cz

Filip Staněk

filip.stanek@vsb.cz

IT4Innovations#
národní01#$%@&0
superpočítačové
centrum$@00&1@&